A global analysis of the relationship between violence and life expectancy

Big Data in Social Sciences

Riccardo Omenti

University of Bologna

Introduction

Measuring **violence** in a country is a challenging task.

- **Violence** is a multifaceted concept
- It depends on many different factors
- Multiple sources are needed to capture its complexities

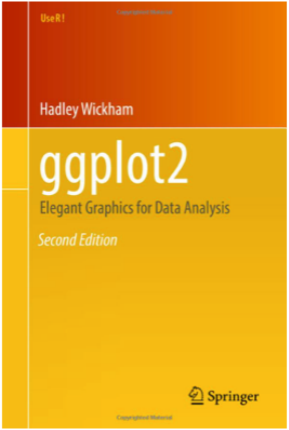
Research Question

- The level of violence may affect the mortality of a country
- Today, we will rely on two major data sources to explore the relationship between violence and mortality:
  - Data on Violece from the Institute for Economics & Peace
  - Data on Mortality from the United Nations
- We will produce multiple descriptive plots with functions from **ggplot**

What about ggplot?

ggplot2 is an R package for producing statistical, or data, graphics

It is already available in *tidyverse*!



Use R!

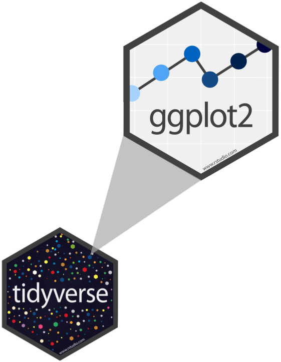
Hadley Wickham

ggplot2

Elegant Graphics for Data Analysis

Second Edition

Springer



ggplot

Major advantages:

- creating graphs by combining independent components
- detailed theming system to generate nice-looking graphs
- intuitive grammar
- graphs are R objects

Indicators for violence and mortality

- **Violence** → **Global Peaceful Index** (GPI) that measures the violence of a country across three dimensions:
  - ongoing domestic and international conflict
  - societal safety and security
  - militarization
- **Mortality** → **life expectancy at birth** (e_0)
  - common indicator for capturing the mortality of a population

Data files

Two *.txt* data files on Virtuale

- *gpi_data.txt* → GPI for multiple countries over 2008-2022
- *life_exp_data.txt* → life expectancy estimates for multiple countries over 2008-2022 couple
  - Regional classification included
  - Total population size

Upload data in Rstudio

Upload *tidyverse* and the data sets

```r
#install.packages('ggthemes')
#install.packages('RColorBrewer')
library('RColorBrewer') # various qualitative color palettes
library('ggthemes') # various themes in ggplot
library("tidyverse") # ggplot2

gpi_data <- read.table('Data/gpi_data.txt',header=T)
life_exp_data <- read.table('Data/life_exp_data.txt',header=T)
```

Combine the two data sets by *iso3* (unique to each country) and *Year*

```r
data <- inner_join(gpi_data,life_exp_data,by=c('iso3','Year'))
```

We match all records in *gpi_data*, whose *iso3* and *Year* values have a correspondence in *life_exp_data*

The non-matching records, either in *gpi_data* or in *life_exp_data*, are dropped

Question 1

Create a plot displaying the evolution of violence over the time period 2008-2022 by world region

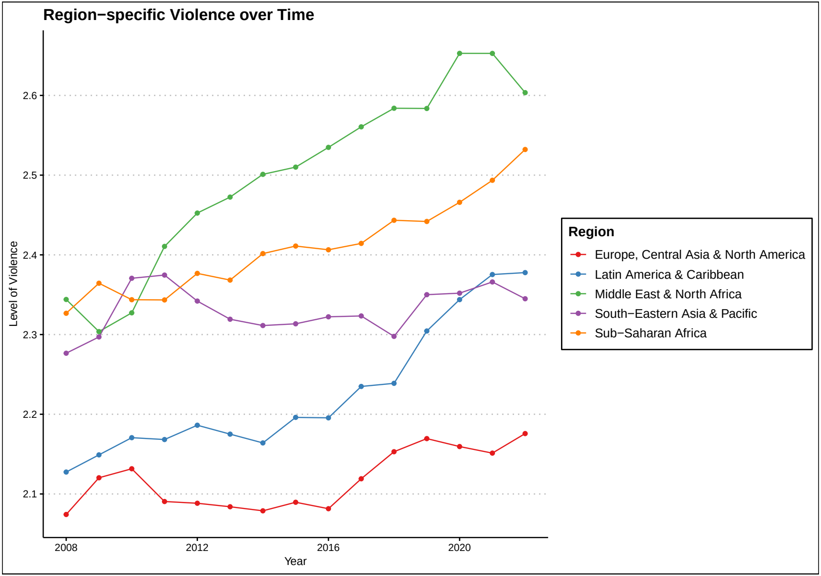
Calculate region- and year-specific levels of violence

```r
data_violence_region <- data |>
      group_by(Year,area) |>
      summarize(GPI=weighted.mean(x=GPI,w=pop,na.rm = FALSE))
```

Generate the plot

```r
plot1 <- ggplot(data=data_violence_region, # data input
     mapping=aes(x=Year,y=GPI,color=area))+ # relationships
geom_line()+ # add a line for each country
geom_point()+ # add points
theme_clean()+ # background
scale_color_brewer(name = "Region", palette = "Set1")+ # color palette
xlab('Year')+ #label for title of x-axis
ylab('Level of Violence') + # label for title of y-axis
ggtitle('Region-specific Violence over Time')  # title for the plot
```

Question 1


Region−specific Violence over Time
Level of Violence
2.1
2.2
2.3
2.4
2.5
2.6
Year
2008
2012
2016
2020
Region
Europe, Central Asia & North America
Latin America & Caribbean
Middle East & North Africa
South−Eastern Asia & Pacific
Sub−Saharan Africa

Question 2

Create a similar visual aid for life expectancy

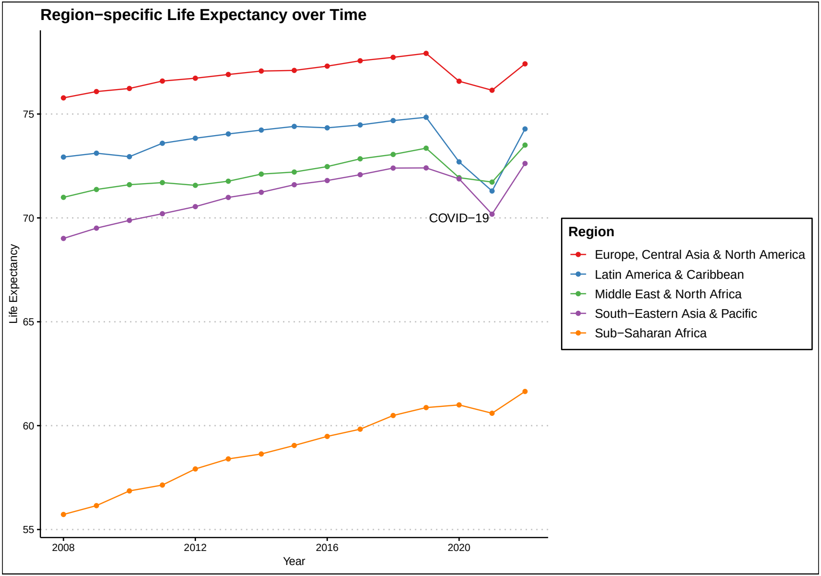
Calculate region-specific life expectancy estimates

```r
data_life_exp_region <- data |>
      group_by(Year,area) |>
      summarize(e0=weighted.mean(x=e0,w=pop,na.rm = FALSE))
```

Generate the plot

```r
plot2 <- ggplot(data=data_life_exp_region, # data input
mapping=aes(x=Year,y=e0,color=area))+ # relationships
geom_line()+ # add a line for each country
geom_point()+ # add points
annotate("text",x=2020,y=70,label="COVID-19")+ # add text
theme_clean()+ # specify a theme
xlab('Year')+ #label for title of x-axis
ylab('Life Expectancy') + # label for title of y-axis
scale_color_brewer(name = "Region", palette = "Set1")+ # change color palette
ggtitle('Region-specific Life Expectancy over Time')  # title
```

Question 2


Region−specific Life Expectancy over Time
Life Expectancy
75
70
65
60
55
2008
2012
2016
2020
Year
COVID−19
Region
Europe, Central Asia & North America
Latin America & Caribbean
Middle East & North Africa
South−Eastern Asia & Pacific
Sub−Saharan Africa

Question 3

Produce two scatter plots to display the relationship between violence and life expectancy in 2008 and 2022. Add also a regression line.

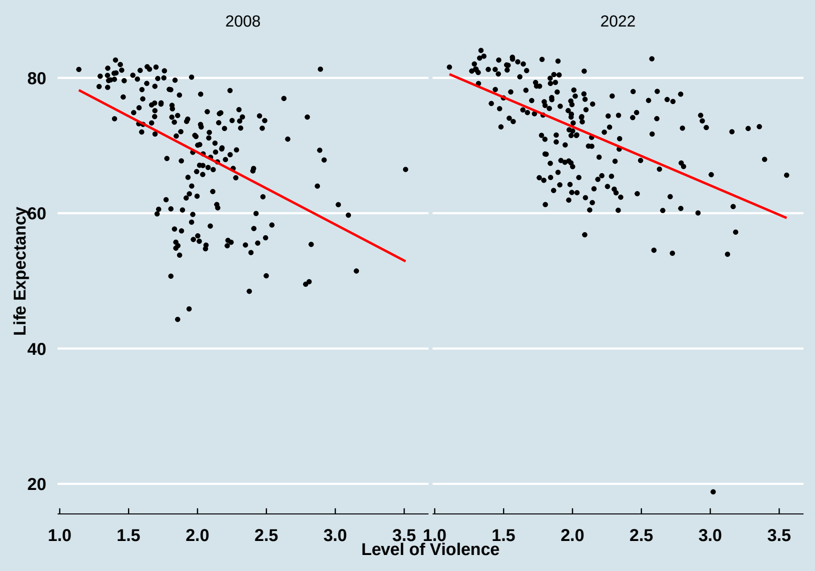
Select records for years 2008 and 2022

```
data_scatter <- data |>
filter(Year %in% c(2008,2022))
```

Generate the plot

```
plot3 <- ggplot(data=data_scatter, # data input
mapping=aes(x=GPI,y=e0))+ # relationships
geom_point()+  # scatter plot
xlab('Level of Violence')+ # x-axis label
ylab('Life Expectancy')+ # y-axis label
theme_economist()+ # pick a background
geom_smooth(method = "lm", se = FALSE,color='red')+ # regression line
facet_wrap(~Year) + # seperate (sub)plot by year
theme(axis.text.y = element_text(size=15,face="bold"), # font of y-axis text
axis.title.y = element_text(size=15,face="bold"), # font of y-axis title
axis.text.x = element_text(size=15,face="bold"),
axis.title.x = element_text(size=15,face="bold")) # font of x-axis title
```

Question 3



2008
2022
Life Expectancy
80
60
40
20
1.0
1.5
2.0
2.5
3.0
3.5
Level of Violence

Question 4

Display the distribution of life expectancy in the 20 most violent countries and in the 20 most peaceful countries in 2022

Let's create the data sets Most violent countries

```r
data_most_violence <- data |>
      filter(Year==2022) |>
      slice_max(GPI,n=20) |>
      mutate(label='Most Violent')
```

Most peaceful countries

```r
data_most_peaceful <- data |>
filter(Year==2022) |>
slice_min(GPI,n=20) |>
mutate(label='Most Peaceful')
```

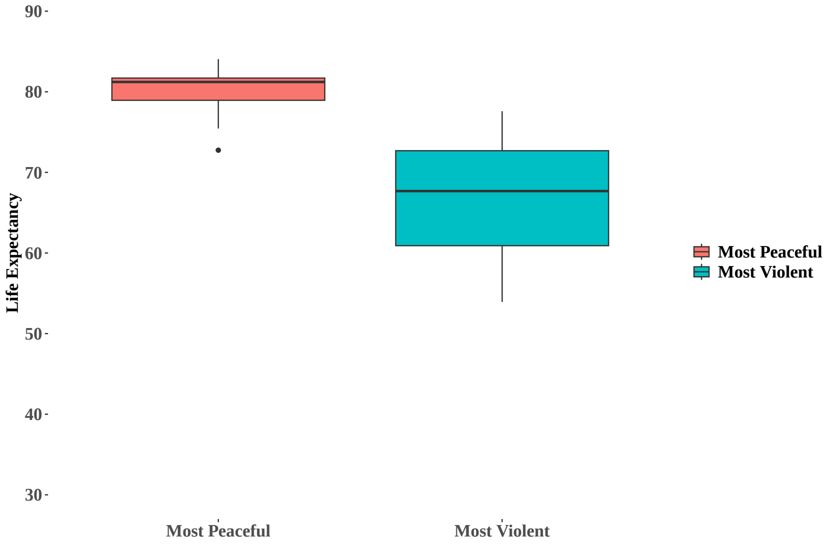
combine the two data sets by row

```r
data_plot <- rbind(data_most_violence,data_most_peaceful)
```

Question 4

```r
plot4 <- ggplot(data=data_plot, # data
mapping=aes(x=label,y=e0,fill=label))+ # relationships
geom_boxplot()+ # boxplot
coord_cartesian(ylim=c(30,90))+ # fix y-axis limits
scale_y_continuous(breaks=seq(30,90,10),
                   labels=seq(30,90,10))+ # fix y-axis labels
theme_tufte()+ # pick background
scale_fill_discrete(name='')+ # drop name from legend
xlab('')+ # no name for x-axis title
ylab('Life Expectancy')+ # y-axis title
theme(axis.text.y = element_text(size=15,face="bold"), # font of y-axis text
legend.text = element_text(size=15,face="bold"), # legend text details
axis.title.y = element_text(size=15,face="bold"), # font of y-axis title
axis.text.x = element_text(size=15,face="bold"),  # text of x-axis title
axis.title.x = element_text(size=15,face="bold")) # font of x-axis title
```

Question 4

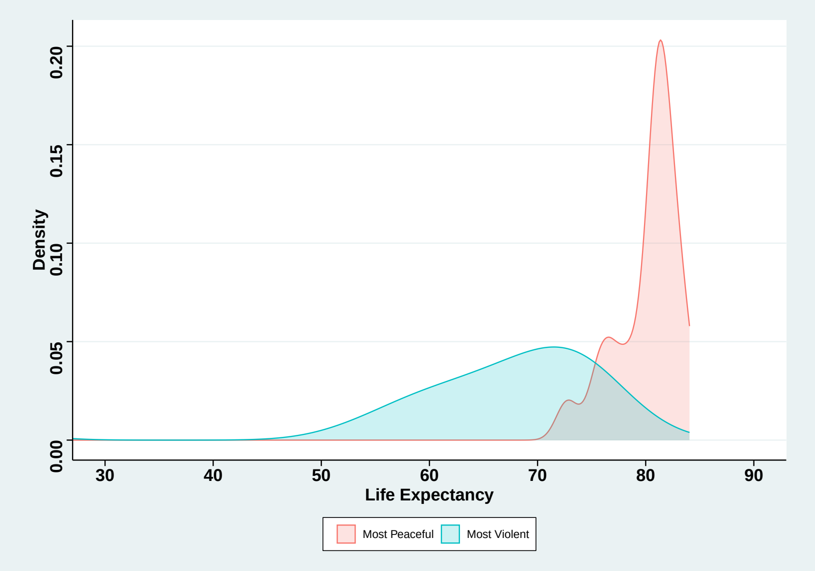

Life Expectancy
90
80
70
60
50
40
30
Most Peaceful
Most Violent
Most Peaceful
Most Violent

Question 5

Perform the same task using a different visual aid.

```r
plot5 <- ggplot(data=data_plot,  # data input
aes(x=e0,color=label,fill=label))+ # relationships
geom_density(alpha = 0.2, na.rm = TRUE) + # density
theme_stata()+ # background
scale_fill_discrete(name='')+ # no title to legend
scale_color_discrete(name='')+ # no title to legend
coord_cartesian(xlim=c(30,90))+ # fix x-axis limits
scale_x_continuous(breaks=seq(30,90,10),
                   labels=seq(30,90,10))+ # control x-axis labels
ylab('Density')+ # y-axis title
xlab('Life Expectancy')+ # x-axis title
theme(axis.text.y = element_text(size=15,face="bold"), # font of y-axis text
legend.position = "bottom", # legend position
axis.title.y = element_text(size=15,face="bold"), # font of y-axis title
axis.text.x = element_text(size=15,face="bold"),  # font of x-axis text
axis.title.x = element_text(size=15,face="bold")) # font of x-axis title
```

Question 5

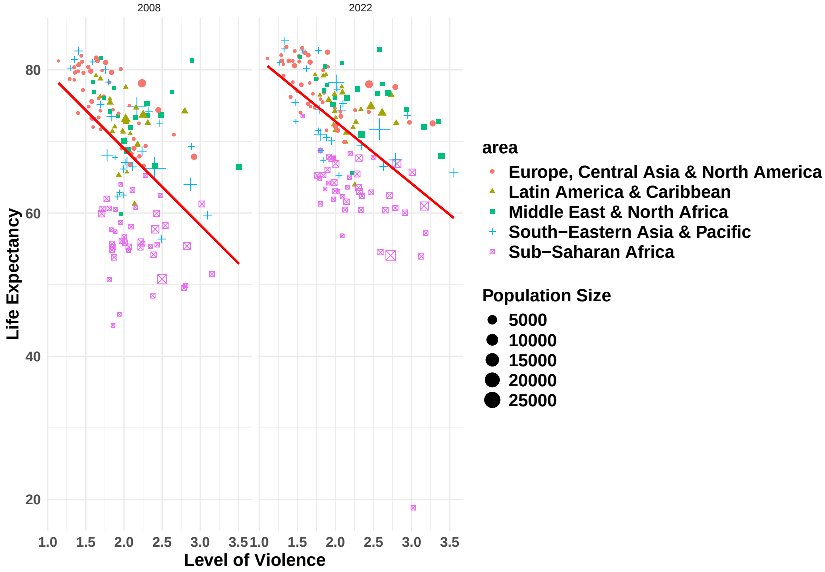

Density
0.00
0.05
0.10
0.15
0.20
30
40
50
60
70
80
90
Life Expectancy
Most Peaceful
Most Violent

Question 6

Produce a scatter plot to display the relationship between violence and life expectancy in 2008 and 2022. Make sure to set different shapes and colors for the points according to the region where they are located. Fix the size of the points according to the population size of the country

```r
plot6 <- ggplot(data=data_scatter, # data input
mapping=aes(x=GPI,y=e0))+ # relationships
geom_point(aes(shape=area,color=area,size=pop))+ # scatter plot
xlab('Level of Violence')+ # label x-axis
ylab('Life Expectancy')+ # label y-axis
theme_minimal()+ # plot them
geom_smooth(method = "lm", se = FALSE,color='red')+ # regression line
labs(size = 'Population Size')+ # legend label
facet_wrap(~Year) + # seperate by year
theme(axis.text.y = element_text(size=12,face="bold"), # font of y-axis text
axis.title.y = element_text(size=15,face="bold"), # font of y-axis title
legend.title = element_text(size=15,face="bold"), # font of legend title
legend.text = element_text(size=15,face="bold"),  # font of legend text
axis.text.x = element_text(size=12,face="bold"),
axis.title.x = element_text(size=15,face="bold")) # font of x-axis title
```

Question 6


2008
2022
Life Expectancy
80
60
40
20
1.0
1.5
2.0
2.5
3.0
3.5
Level of Violence
area
Europe, Central Asia & North America
Latin America & Caribbean
Middle East & North Africa
South−Eastern Asia & Pacific
Sub−Saharan Africa
Population Size
5000
10000
15000
20000
25000

Question 7

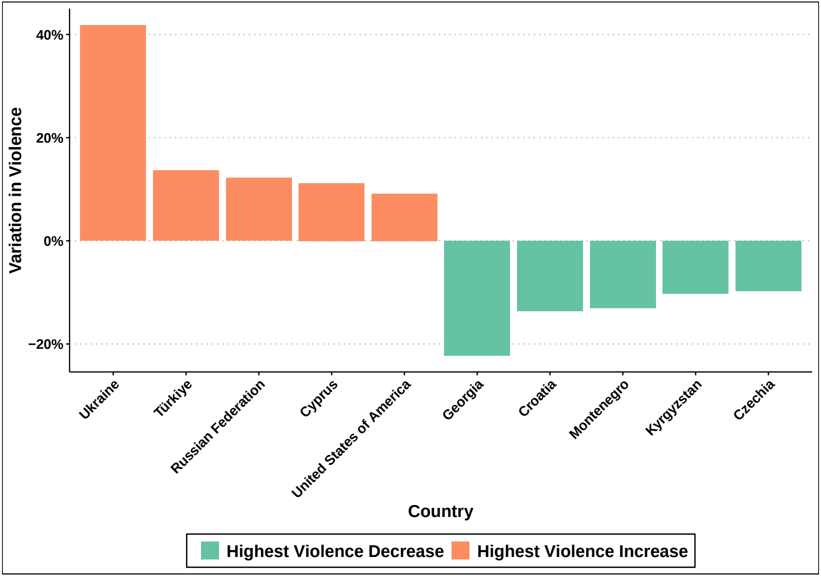
Display the variation in the level of violence between 2008 and 2022 for the top 5 countries experiencing the largest increase and the top 5 showing the largest decrease in violence in Europe, Central Asia & North America

```r
data_high <- data |>
filter(Year %in% c(2008,2022)) |>
select(Year,Country,area,GPI) |>
pivot_wider(names_from='Year',values_from='GPI') |>
mutate(var=(`2022`-`2008`)/`2008`) |>
group_by(area)|>
slice_max(var,n=5) |>
mutate(label='Highest Violence Increase')
data_low <- data |>
filter(Year %in% c(2008,2022)) |>
select(Year,Country,area,GPI) |>
pivot_wider(names_from='Year',values_from='GPI') |>
mutate(var=(`2022`-`2008`)/`2008`) |>
group_by(area)|>
slice_min(var,n=5) |>
mutate(label='Highest Violence Decrease')
```

Question 7

```
plot7 <- rbind(data_high,data_low) |>
  filter(area=='Europe, Central Asia & North America') |>
  mutate(Country=factor(Country,levels=c(data_high$Country,
                                         data_low$Country))) |>
  ggplot(aes(x=Country,y=var,fill=label))+ # data
  geom_bar(stat='identity') + # bar plot
  theme_clean()+ # background
  xlab('Country')+ # x-axis title
  ylab('Variation in Violence')+ # y-axis title
  scale_fill_brewer(name='',palette='Set2')+ # color palette
  scale_y_continuous(labels = scales::percent)+ # y-axis labels
  theme(axis.text.x = element_text(angle = 45,
  vjust = 1, hjust = 1, size = 12, face = "bold"), # x-axis text
        legend.position = 'bottom', # legend position
        axis.text.y = element_text(size = 12, face = "bold"), # y-axis text
        axis.title.x = element_text(size = 15, face = "bold"), # x-axis title
        axis.title.y = element_text(size = 15, face = "bold"), # y-axis title
        legend.text = element_text(size = 15, face = "bold")) # legend text
```

Question 7



Variation in Violence
40%
20%
0%
−20%
Ukraine
Türkiye
Russian Federation
Cyprus
United States of America
Georgia
Croatia
Montenegro
Kyrgyzstan
Czechia
Country
🟩 Highest Violence Decrease
🟧 Highest Violence Increase

Save a plot

ggsave() allows to save plots as images in different formats (e.g. .png, .pdf, .jpeg)

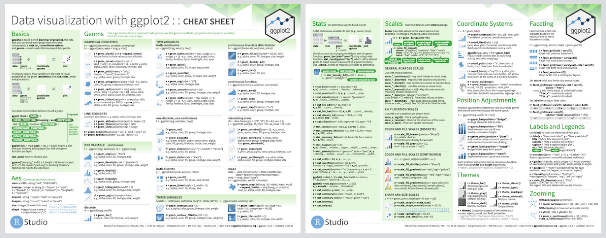
```r
ggsave(filename = "Results/plot7.pdf", # name of the file
       plot = plot7, # name of the plot
       height = 20,  # height of the figure
       width = 40, # width of the figure
       units = "cm", # unit of measurement
       dpi = 400) # resolution of picture
```

Note that graphs can also be saved by pressing the Export bottom in the *Plots* section

Final remarks

What I presented today is just the tip of the iceberg.

For other exciting visualizations see the online book https://ggplot2-book.org



Data visualization with ggplot2 : : CHEAT SHEET

Basics

Geoms

Aes

Stats

Scales

Coordinate Systems

Position Adjustments

Themes

Faceting

Labels and Legends

Zooming